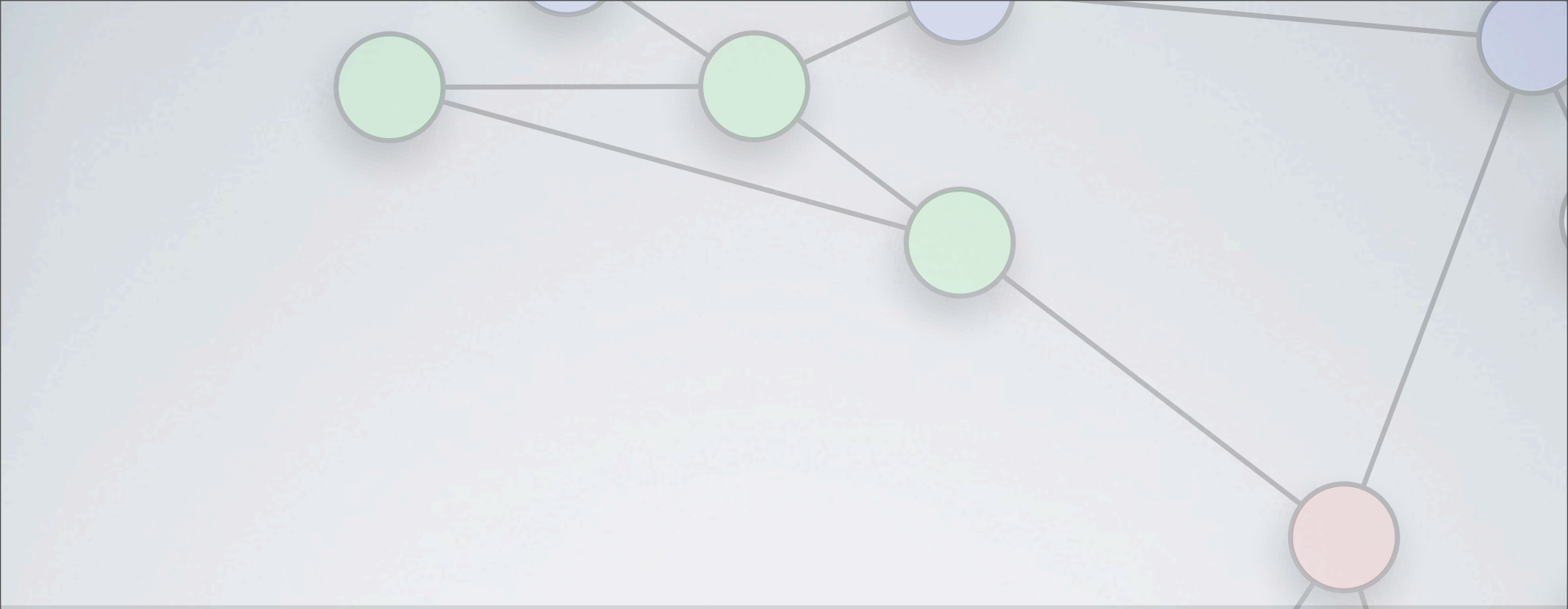


a generative model of hierarchical structure

aaron clauset
computer science dept. & biofrontiers institute
university of colorado, boulder
external faculty, santa fe institute

SIAM Review
Vol. 51, No. 4, pp. 661–703

Power-Law Distributions in Empirical Data*

Aaron Clauset[†]
Cosma Rohilla Shalizi[‡]
M. E. J. Newman[§]

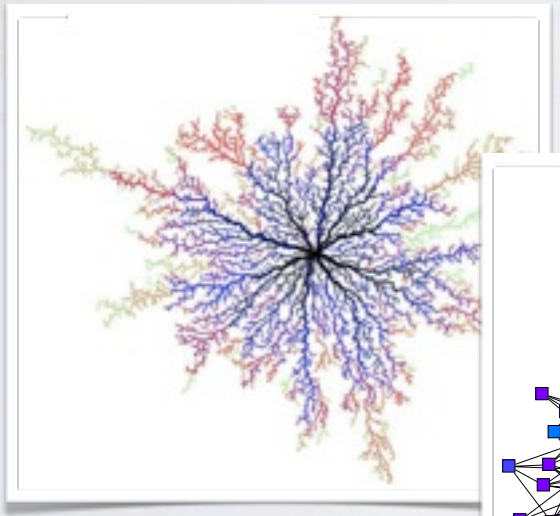
Abstract. Power-law distributions occur in many situations of scientific interest and have significant consequences for our understanding of natural and man-made phenomena. Unfortunately, the detection and characterization of power laws is complicated by the large fluctuations that occur in the tail of the distribution—the part of the distribution representing large but rare events—and by the difficulty of identifying the range over which power-law behavior holds. Commonly used methods for analyzing power-law data, such as least-squares fitting, can produce substantially inaccurate estimates of parameters for power-law distributions, and even in cases where such methods return accurate answers they are still unsatisfactory because they give no indication of whether the data obey a power law at all. Here we present a principled statistical framework for discerning and quantifying power-law behavior in empirical data. Our approach combines maximum-likelihood fitting methods with goodness-of-fit tests based on the Kolmogorov–Smirnov (KS) statistic and likelihood ratios. We evaluate the effectiveness of the approach with tests on synthetic data and give critical comparisons to previous approaches. We also apply the proposed methods to twenty-four real-world data sets from a range of different disciplines, each of which has been conjectured to follow a power-law distribution. In some cases we find these conjectures to be consistent with the data, while in others the power law is ruled out.

1. fit your data
2. validate the model
3. compare to alternatives

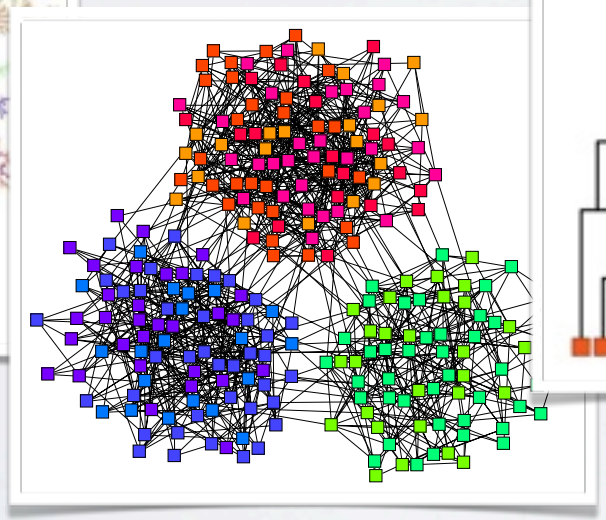
santafe.edu/~aaronc/powerlaws/

large-scale structures

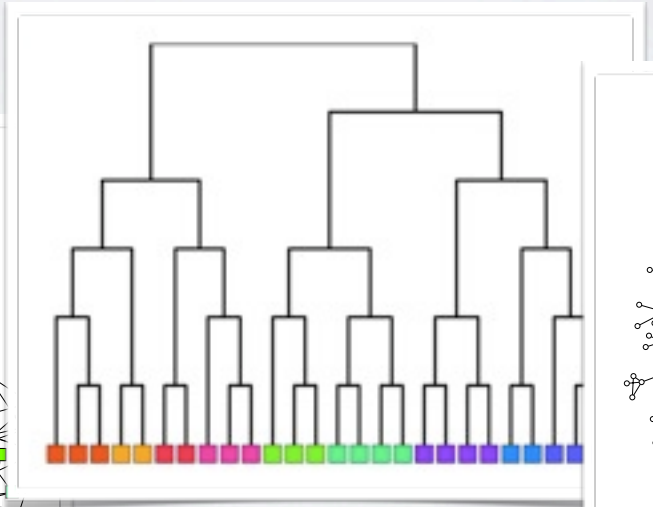
- global patterns: spatial structure, modules, hierarchies, etc.
- what impact on system function?
- what micro-processes build them?
- how can we detect them?



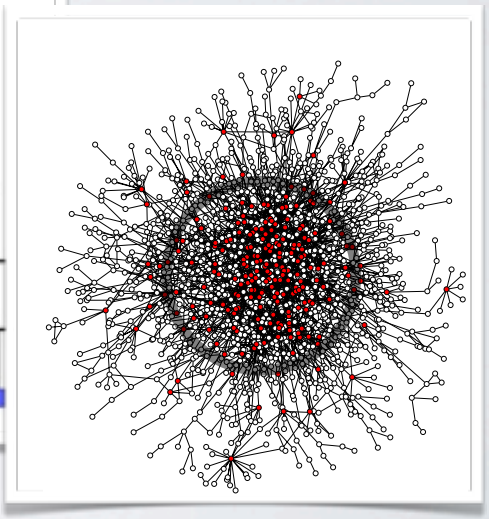
spatial



modular



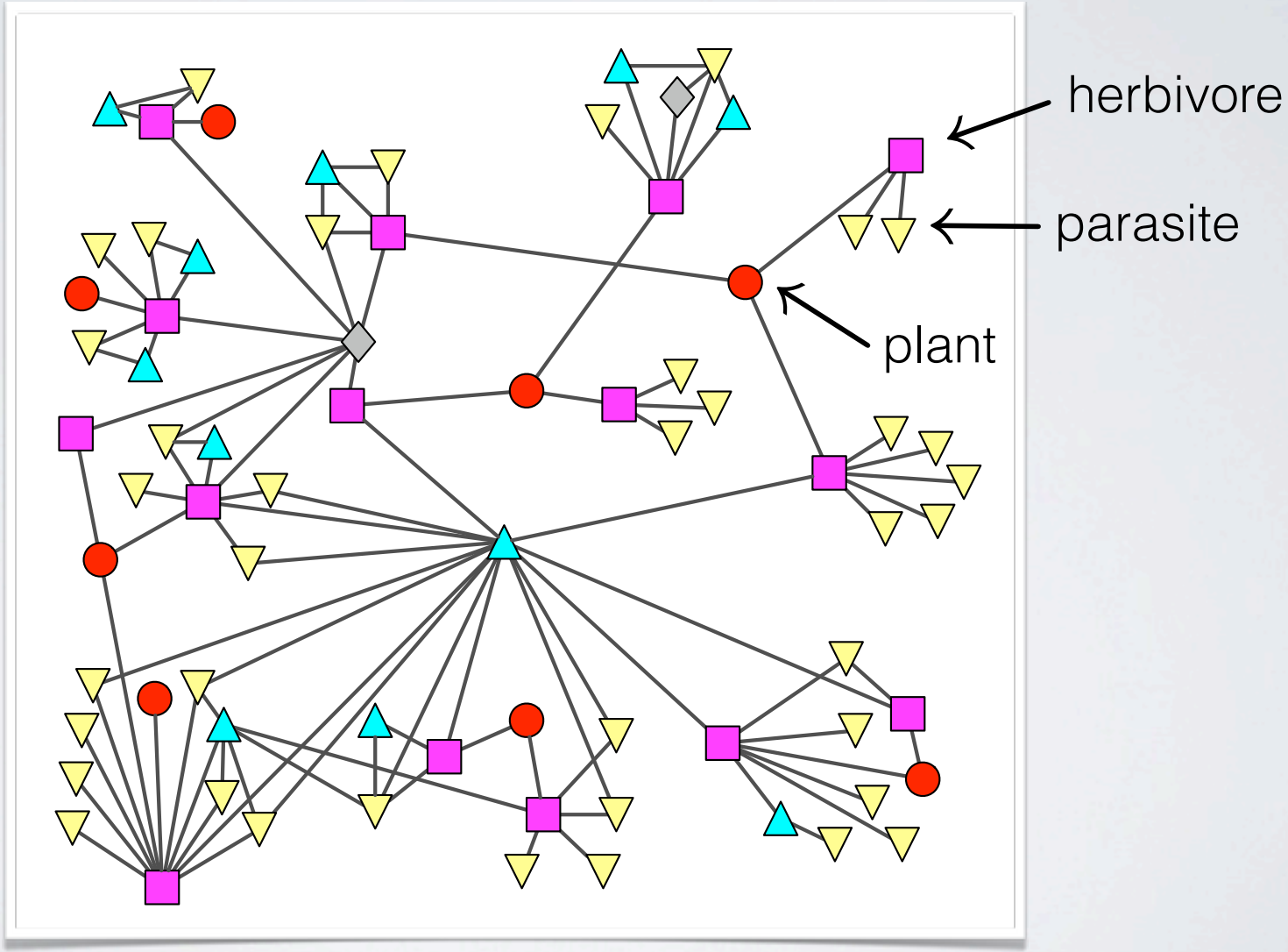
hierarchical



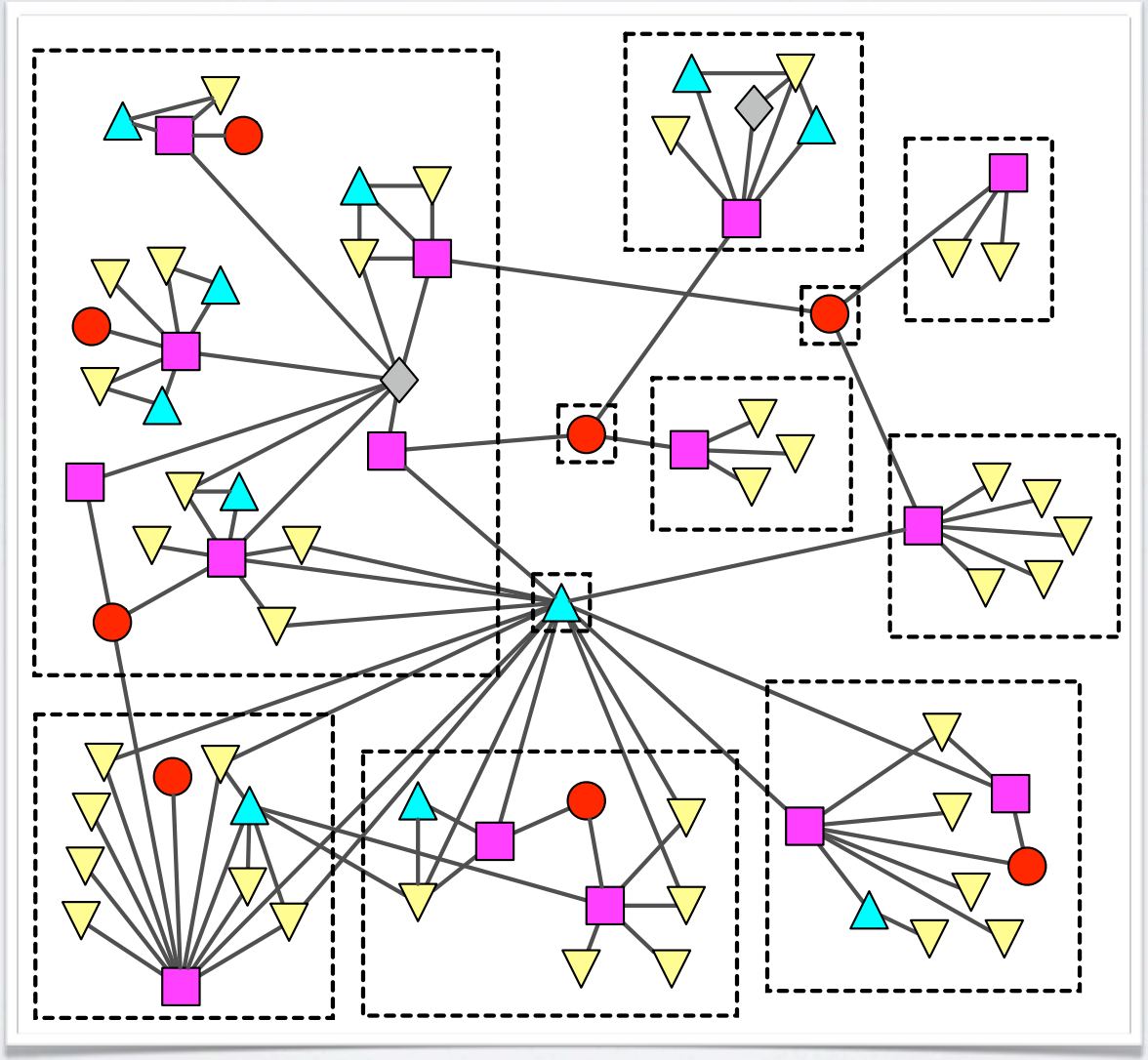
core-periphery

hierarchical structure

hierarchical structure

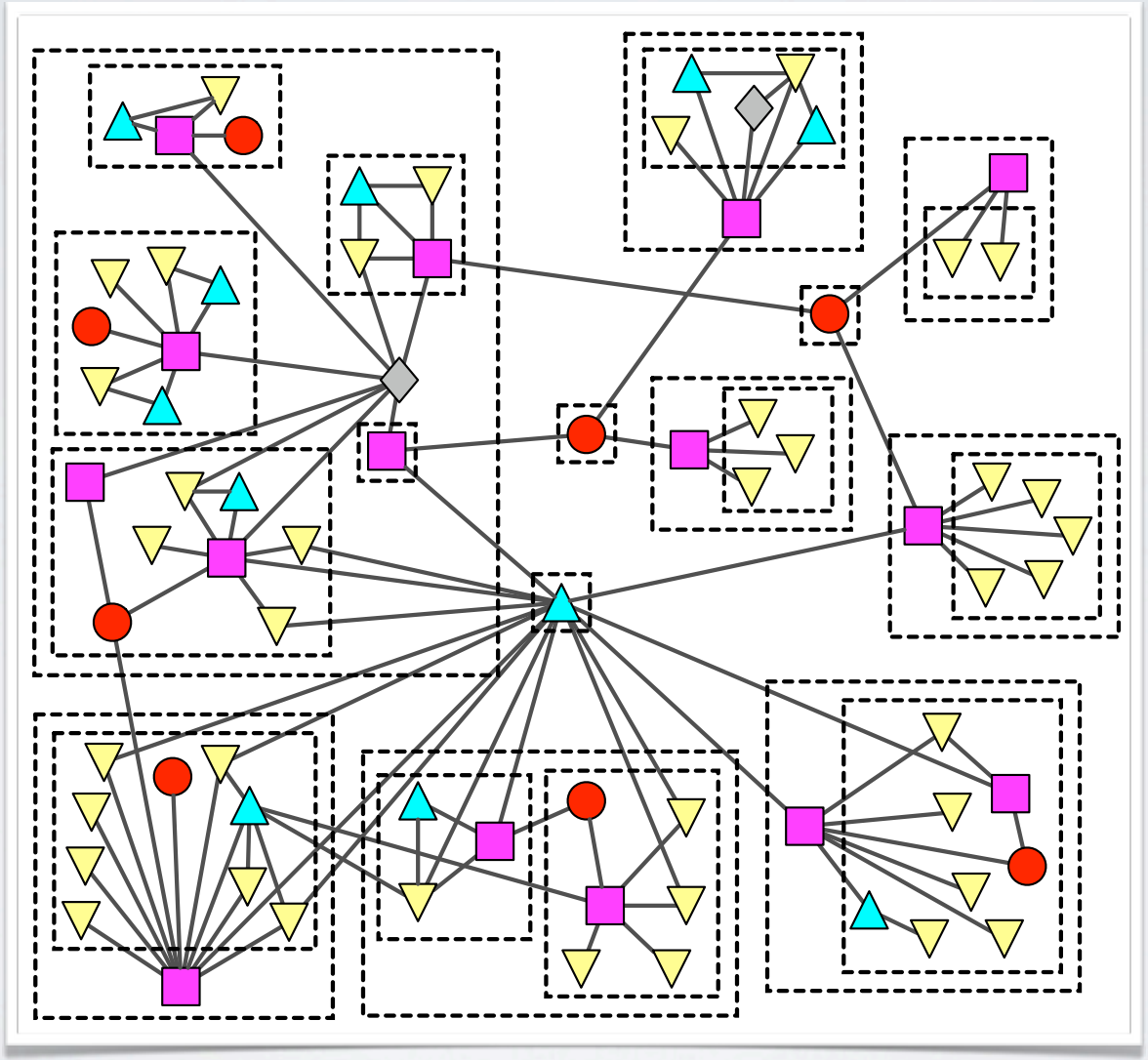


hierarchical structure



modules

hierarchical structure

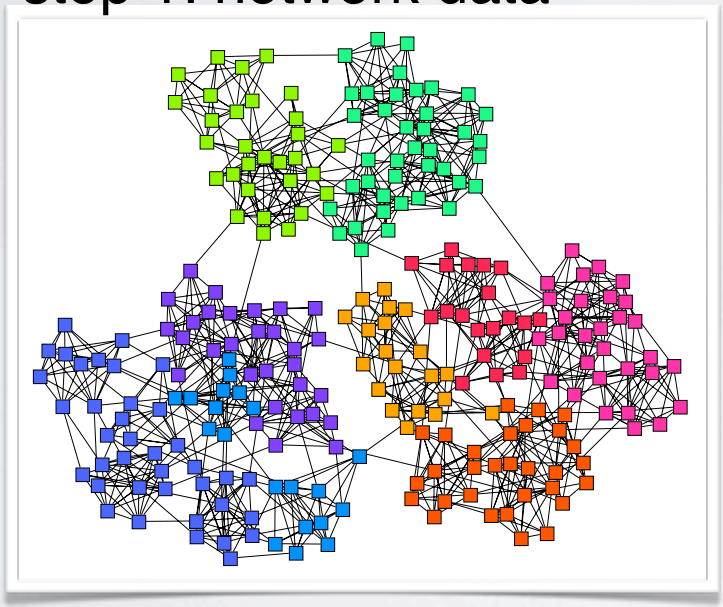


nested modules

hierarchical structure

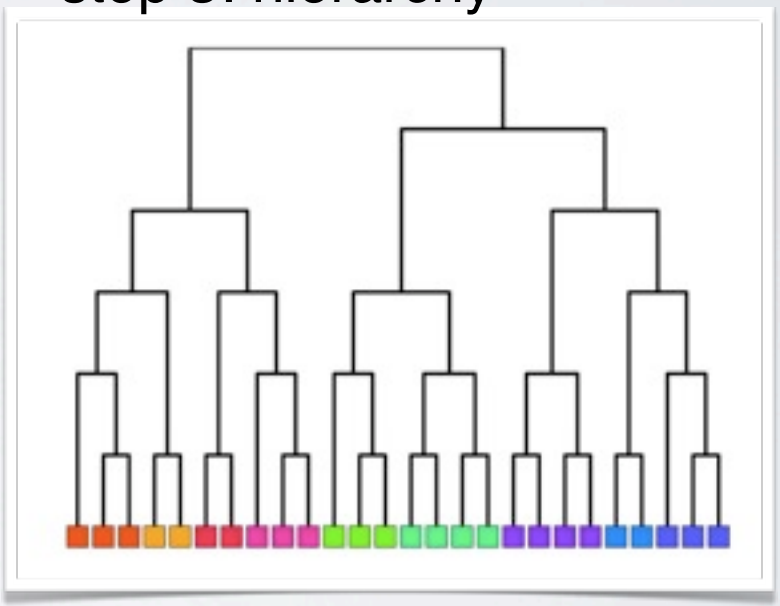
can we automatically extract hierarchies?

step 1: network data



?

step 3: hierarchy



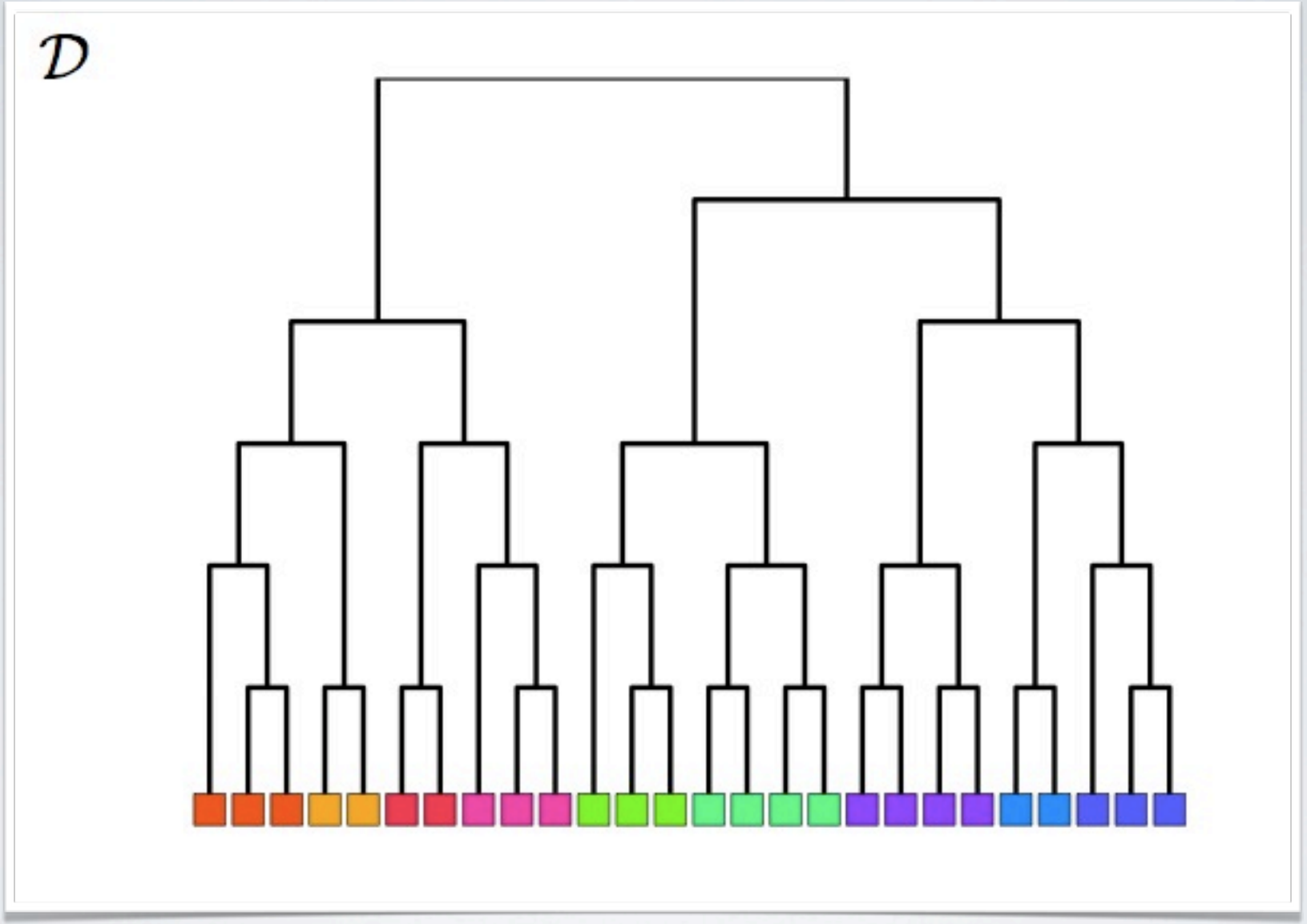
Clauset, Moore, Newman, ***Nature* 453**, 98-101 (2008)
Clauset, Moore, Newman, *ICML* (2006)

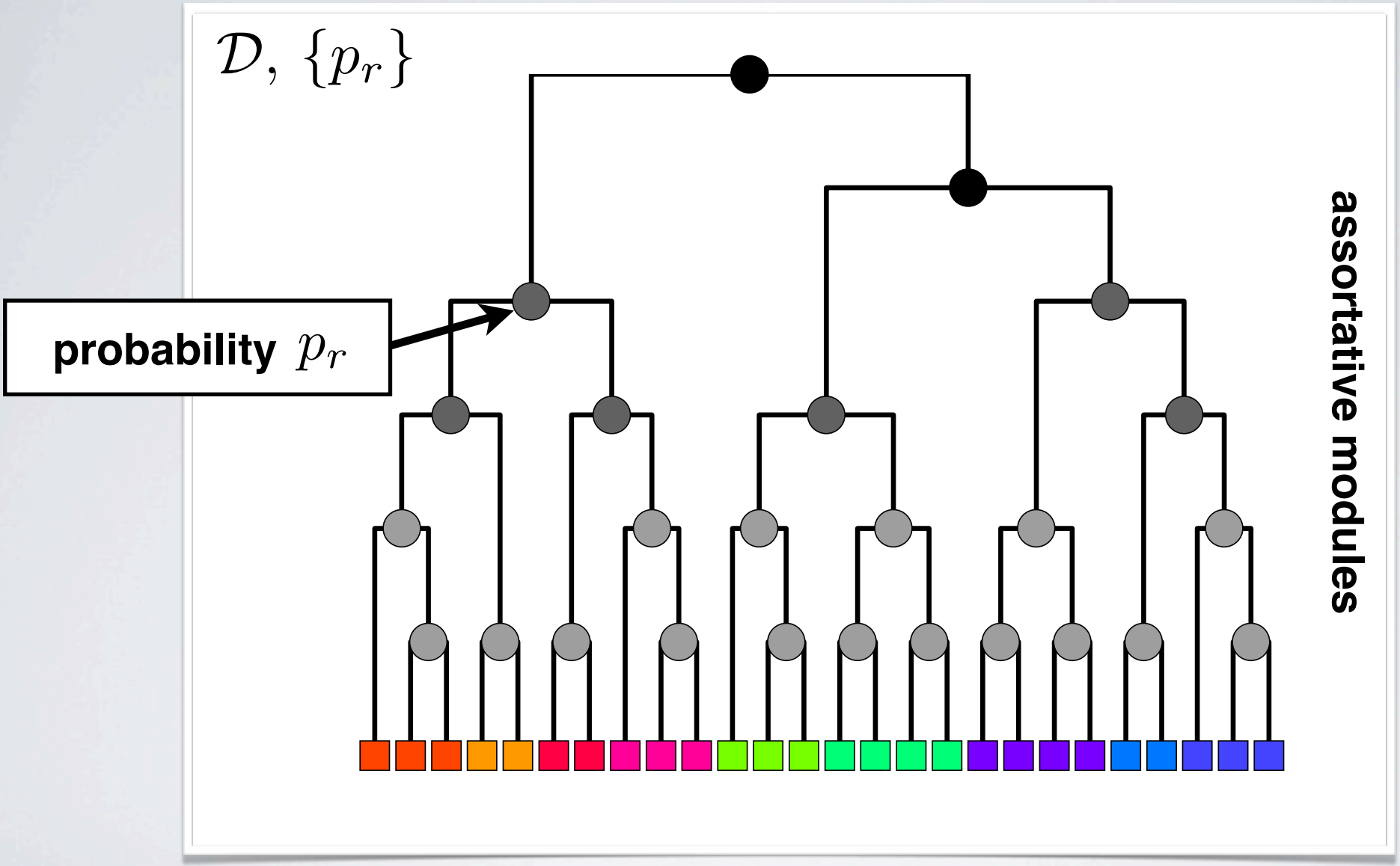
learning from network data

a direct approach

- **write down (edge) generative model** $\Pr(G \mid \theta)$
- **sample models** via MCMC* $\rightarrow \{\theta_i\}$
- **use ensemble** $\{\theta_i\}$ **to test fit, make predictions**
- technical details in *Nature* (2008) and *ICML* (2006)

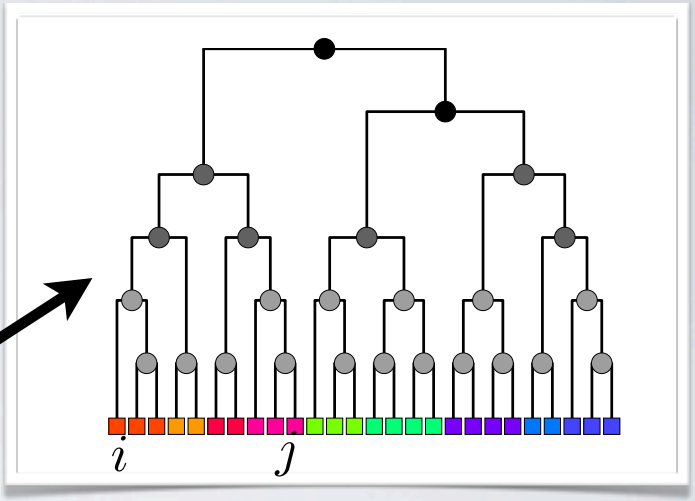
* other sampling or optimization methods possible

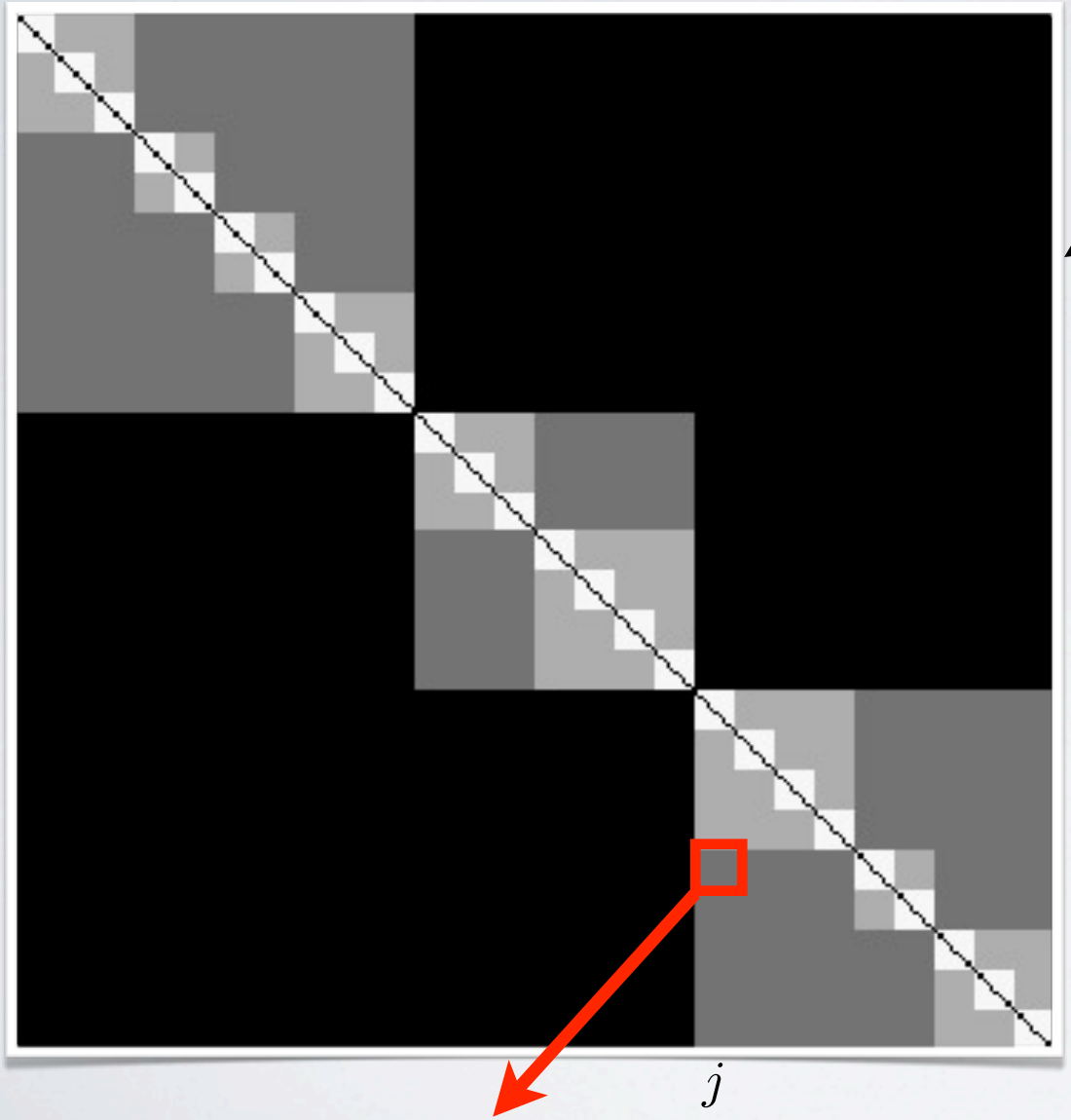

$\mathcal{D}$


$\mathcal{D}$, $\{p_r\}$
probability p_r
assortative modules

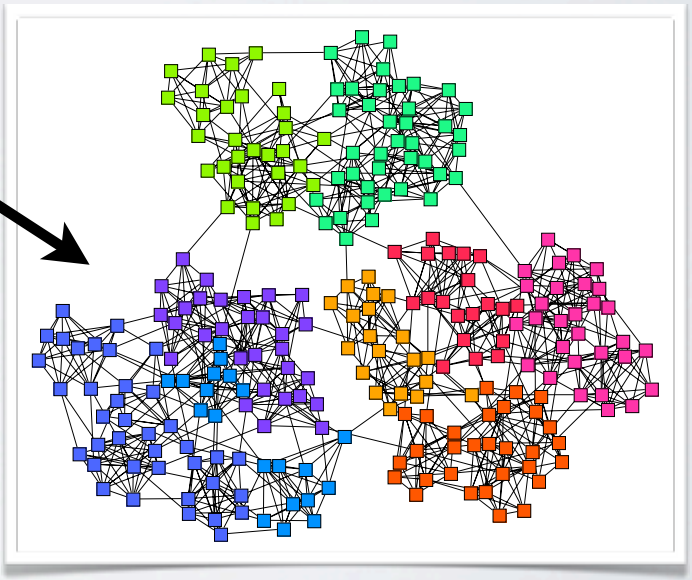
model

“inhomogeneous” random graph





instance



$$\Pr(i,j \text{ connected}) = p_r$$
$$= p_{(\text{lowest common ancestor of } i,j)}$$

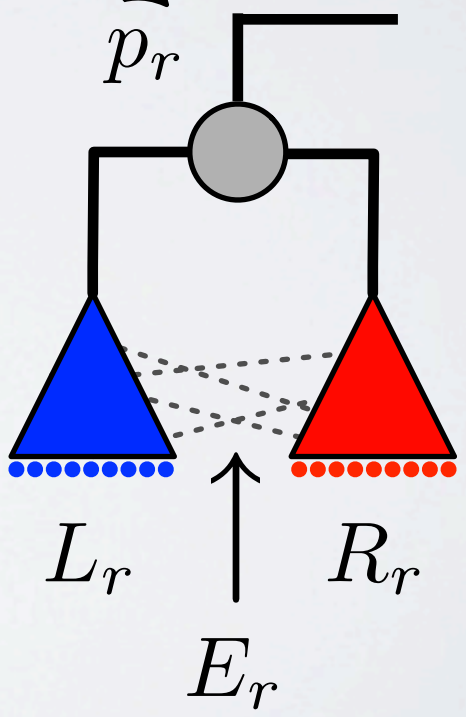
the likelihood function

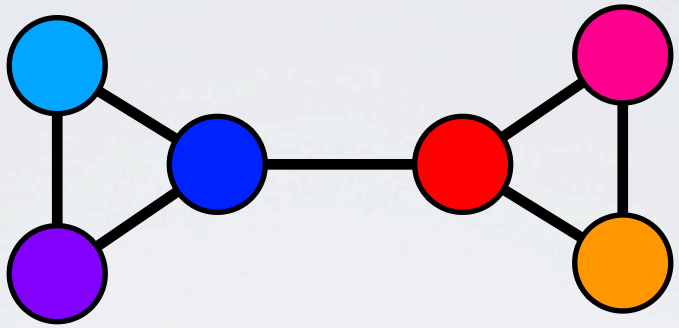
$$\mathcal{L}(\mathcal{D}, \{p_r\}) = \prod_r p_r^{E_r} (1 - p_r)^{L_r R_r - E_r}$$

L_r = number nodes in left subtree

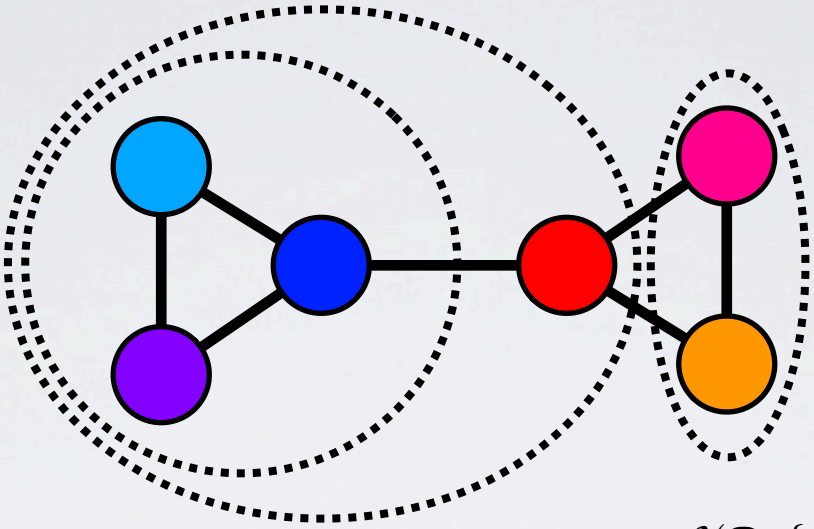
R_r = number nodes in right subtree

E_r = number edges with r as lowest common ancestor

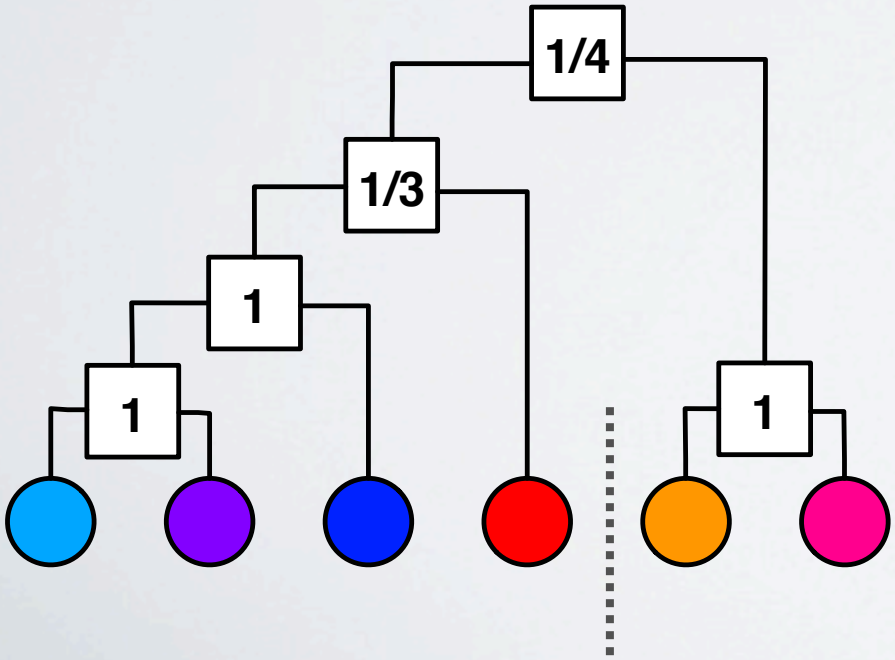




bad dendrogram



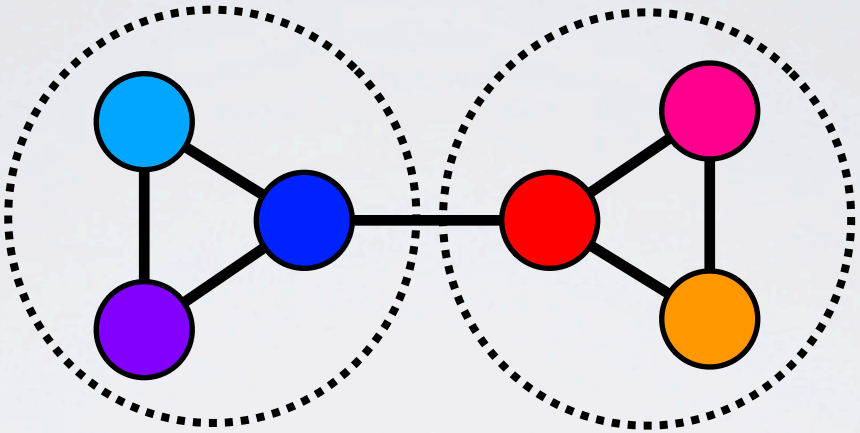
$$\mathcal{L}(\mathcal{D}, \{p_r\}) = \prod_r p_r^{E_r} \left(1 - p_r\right)^{L_r R_r - E_r}$$



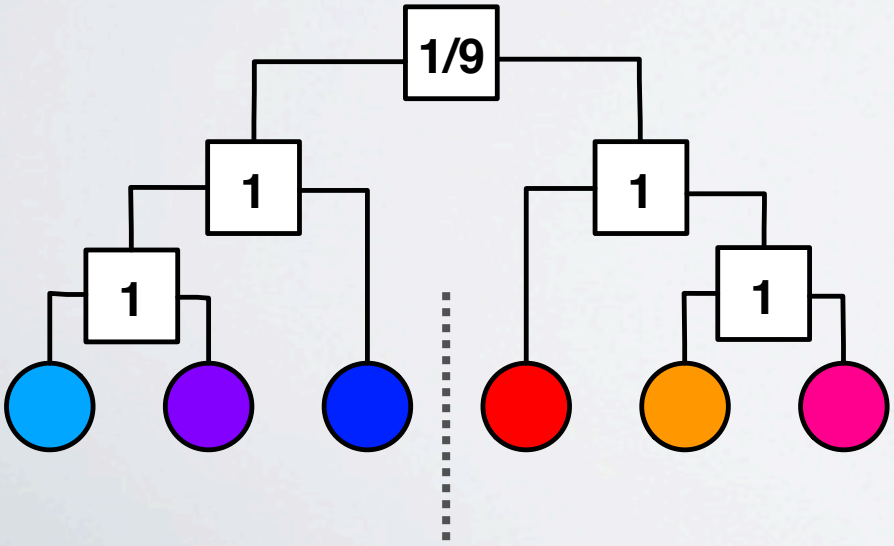
$$\mathcal{L} = \left[\left(\frac{1}{3}\right)^1 \left(\frac{2}{3}\right)^2\right] \cdot \left[\left(\frac{1}{4}\right)^2 \left(\frac{3}{4}\right)^6\right]$$

$$\mathcal{L} = 0.0016$$

good dendrogram



$$\mathcal{L}(\mathcal{D}, \{p_r\}) = \prod_r p_r^{E_r} \left(1 - p_r\right)^{L_r R_r - E_r}$$



$$\mathcal{L} = \left[\left(\frac{1}{9}\right)^1 \left(\frac{8}{9}\right)^8 \right]$$

$$\mathcal{L} = 0.0433$$

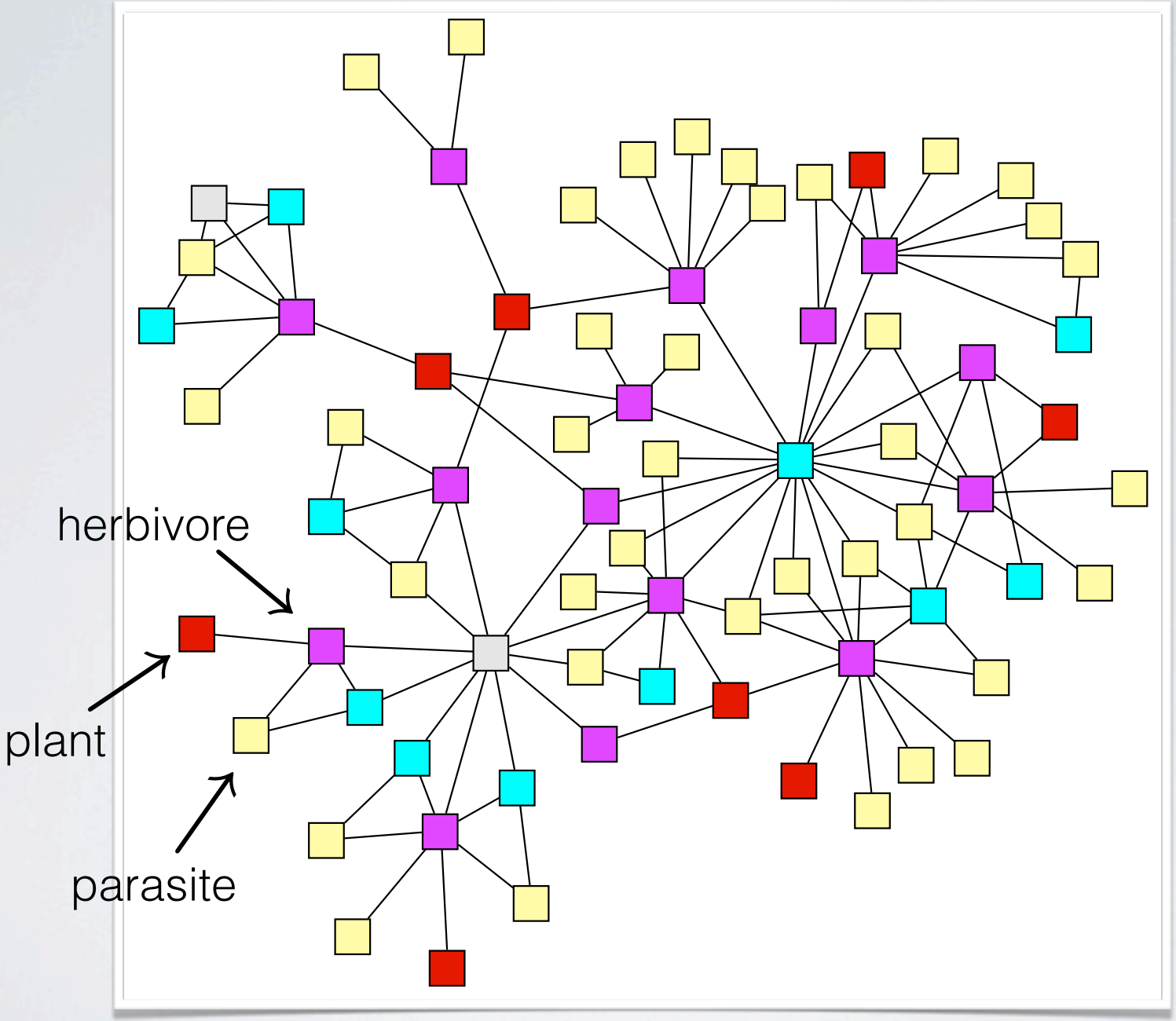
generalizing from a single example

generalizing from a single example

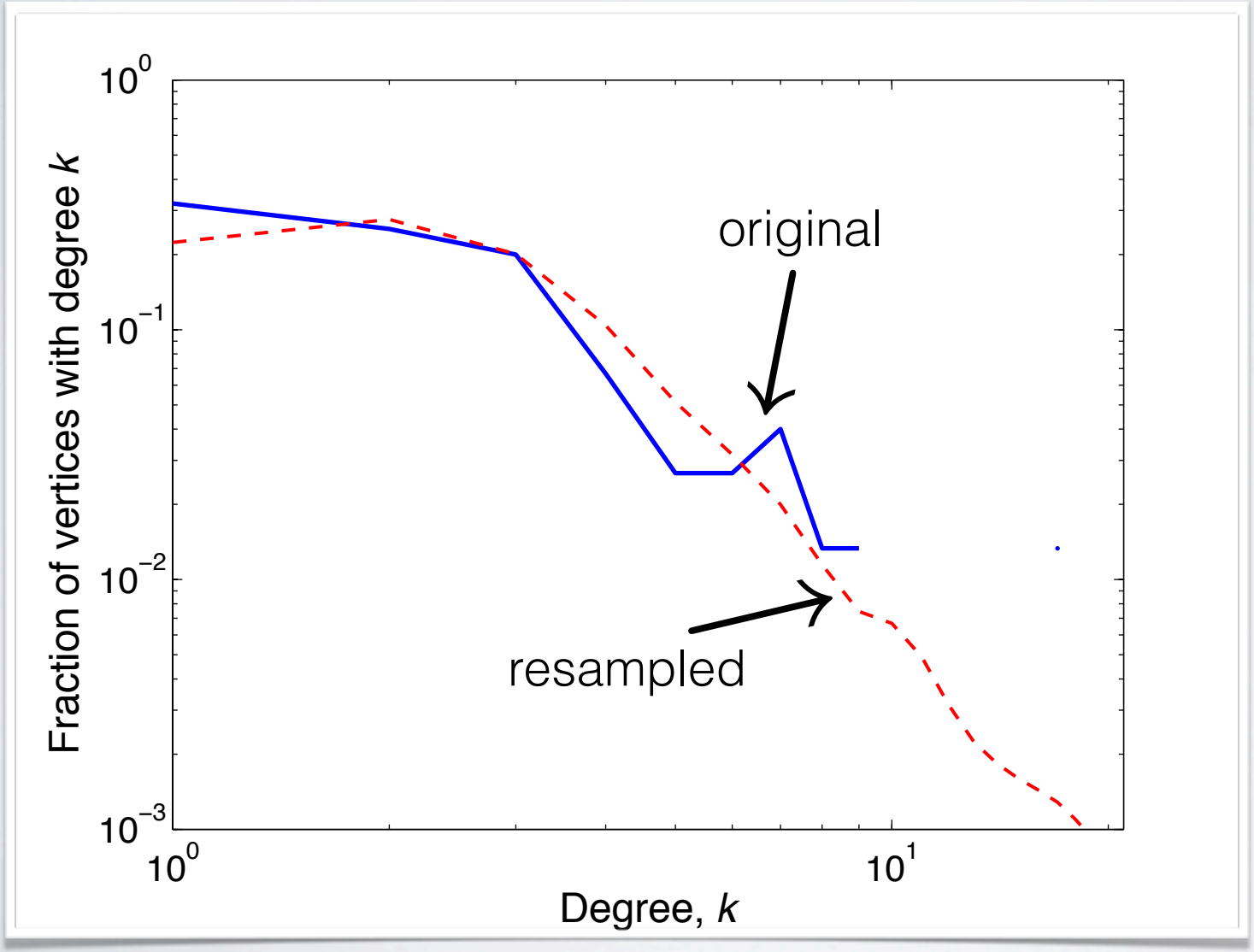
- given graph G
- run MCMC to equilibrium
- for each sampled $\mathcal{D}$, draw a new graph G' from ensemble

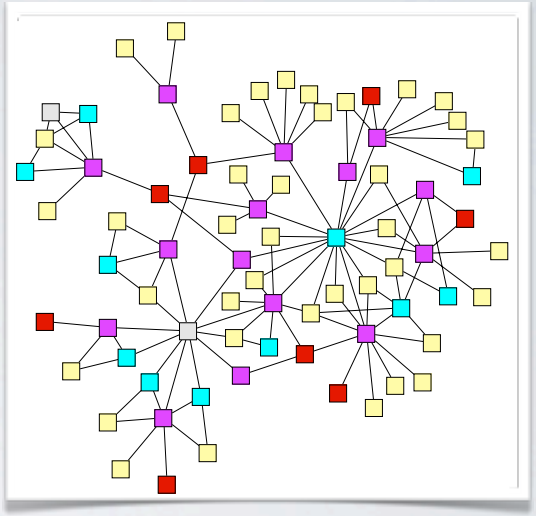
test of model fit:

do resampled graphs look like original?

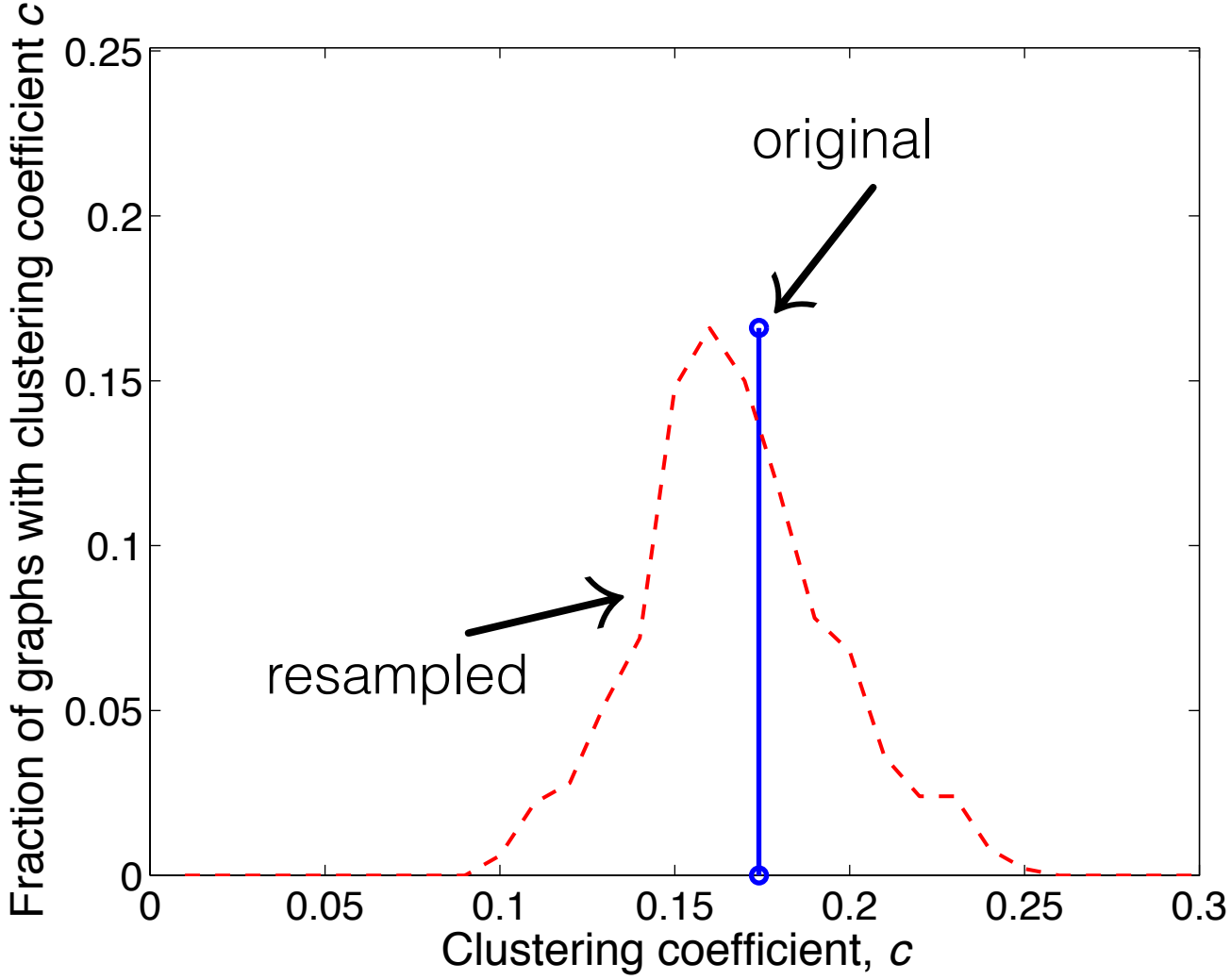

herbivore
plant
parasite

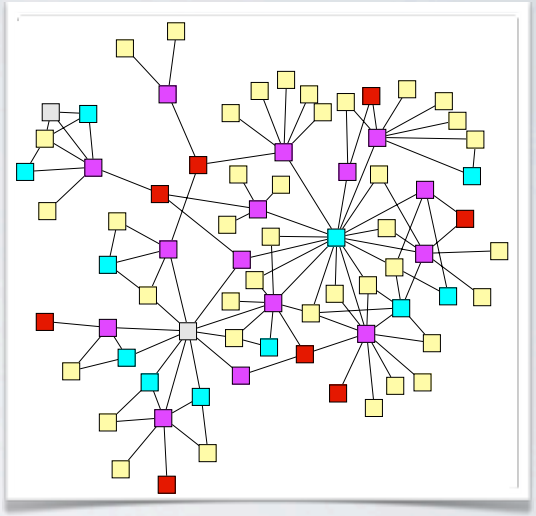
degree distribution



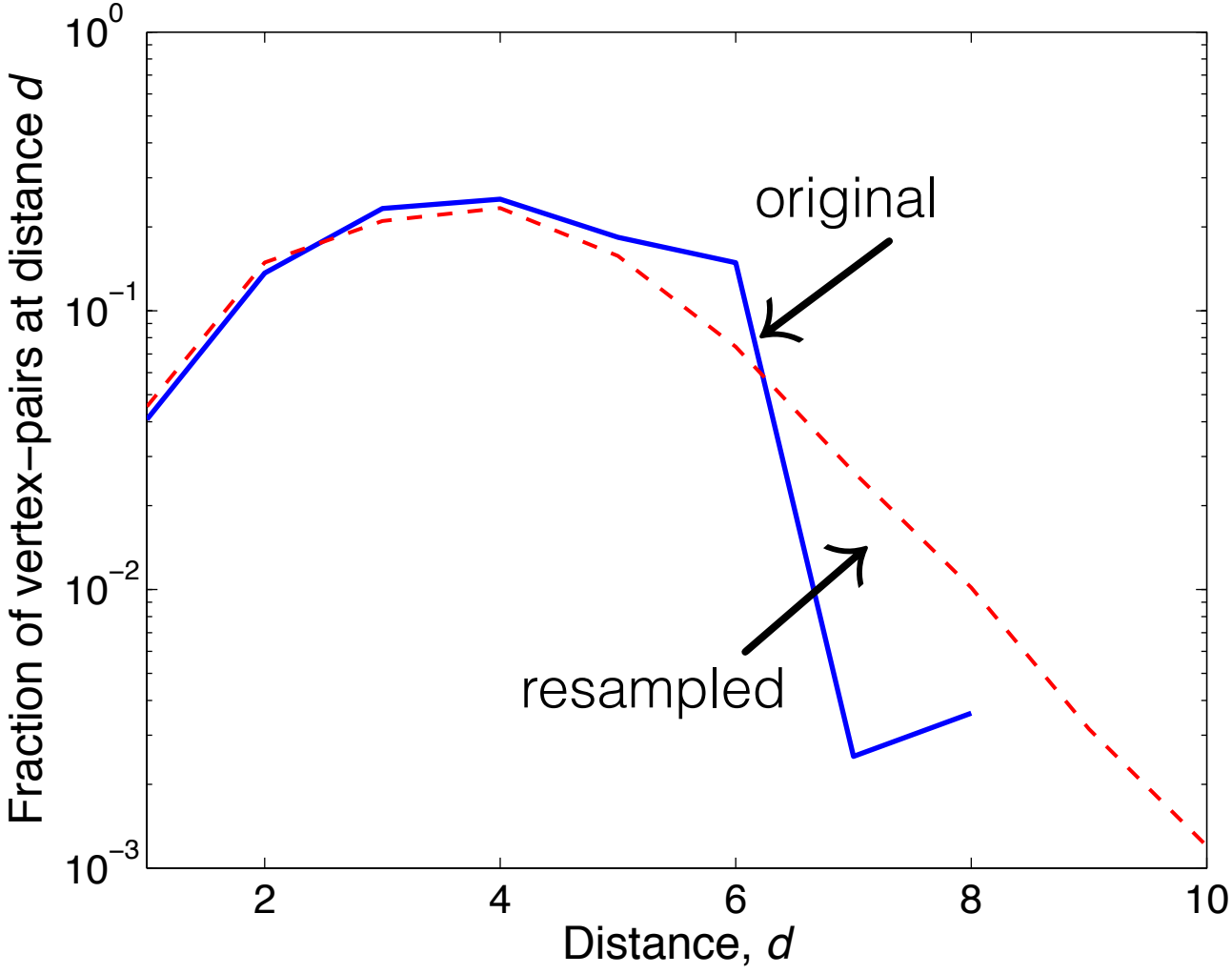


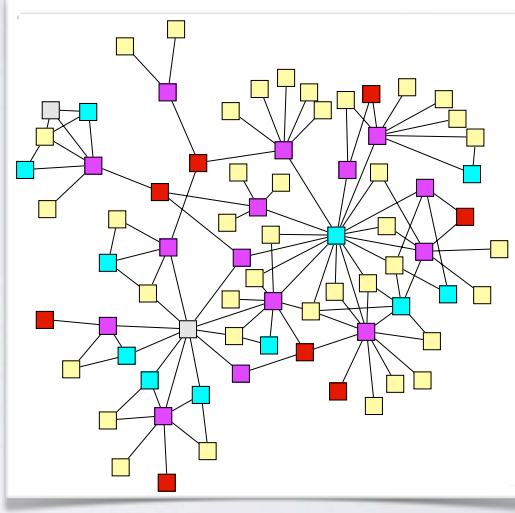
clustering coefficient





distance distribution





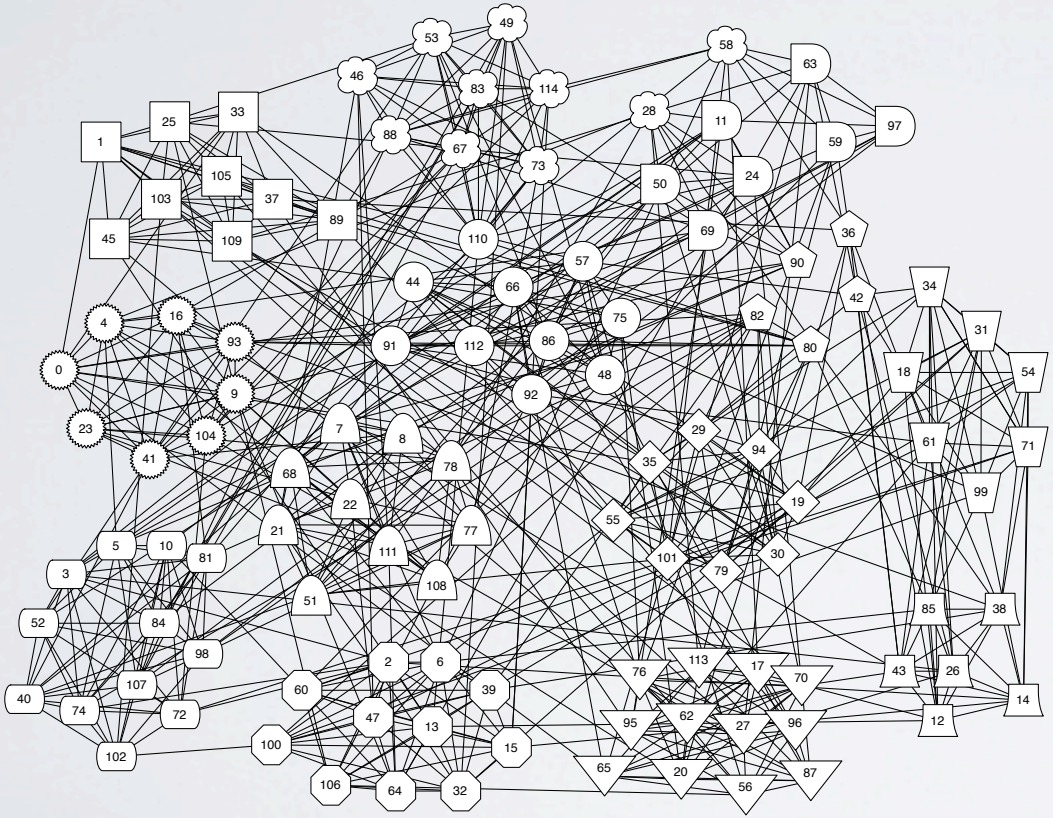
consensus hierarchies

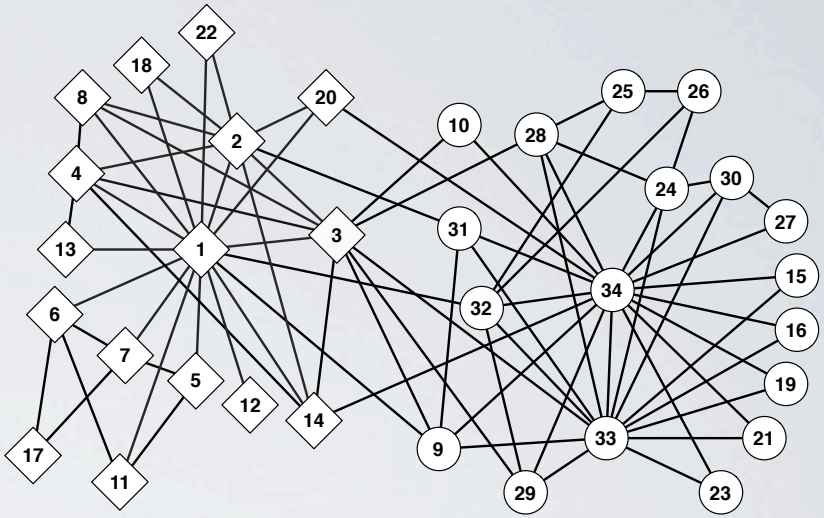
phylogenetic tree reconstruction $\rightarrow$ extract "consensus" tree

yields set of hierarchical relationships $\mathcal{D}_c$ contained in majority of sampled hierarchies $\{\mathcal{D}_i\}$

NCAA Schedule 2000

$n = 115 \quad m = 613$





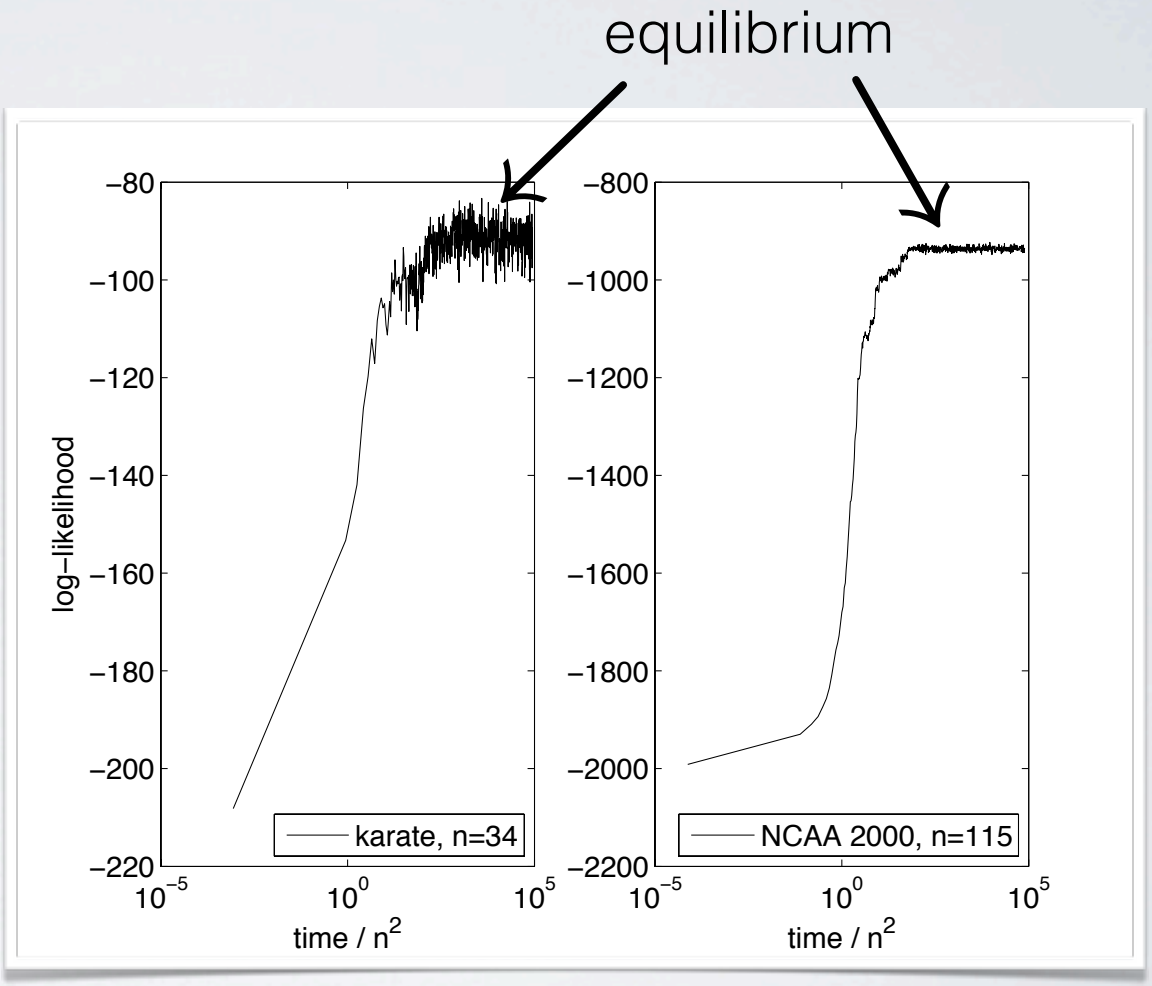
Zachary's Karate Club

$n = 34 \quad m = 78$

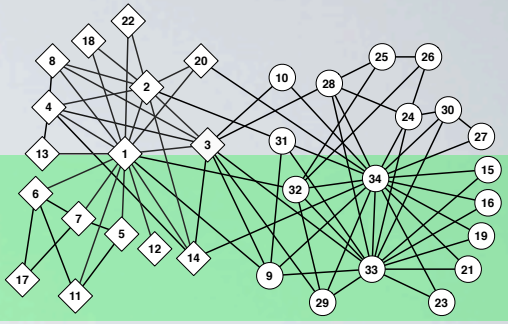
mixing times

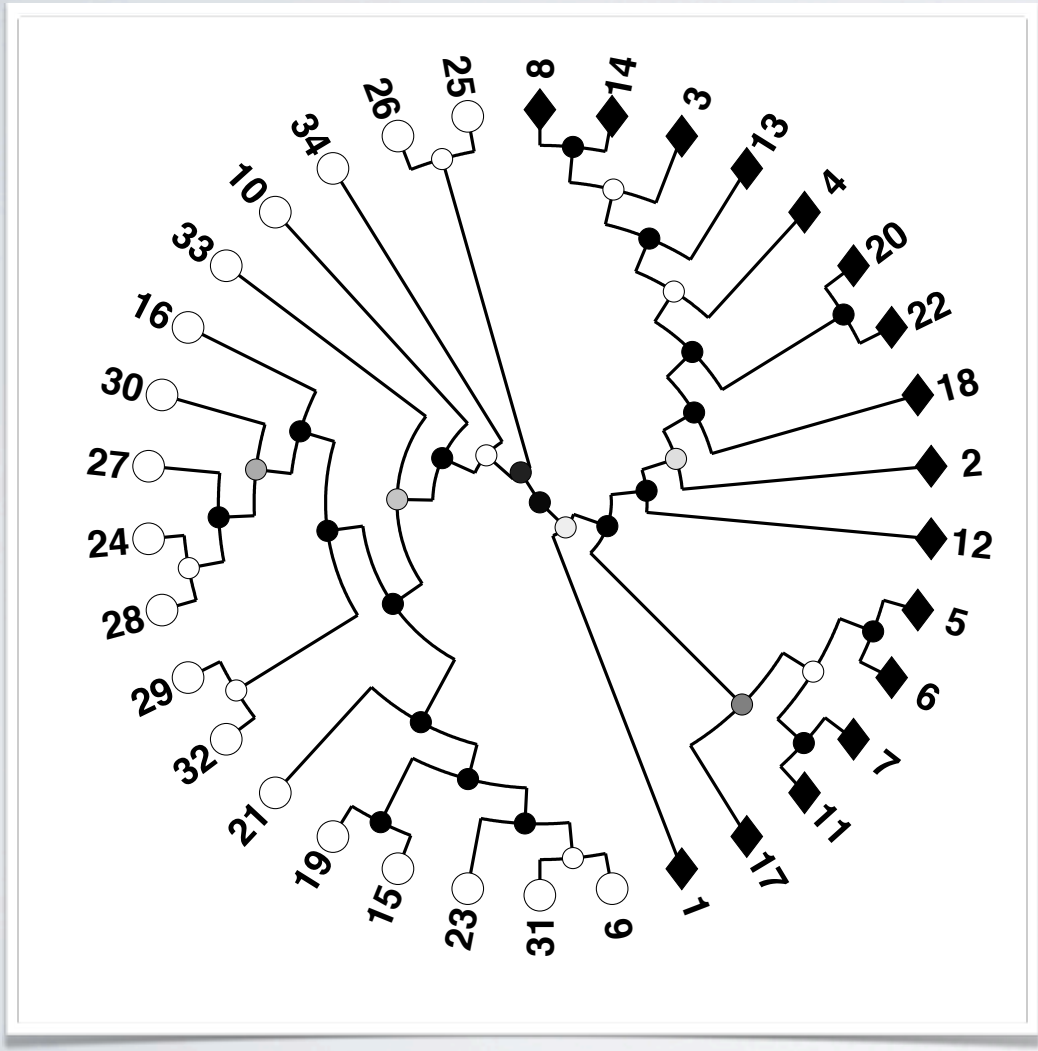
MCMC mixes relatively quickly

Equilibrium in $O(n^2)$ steps

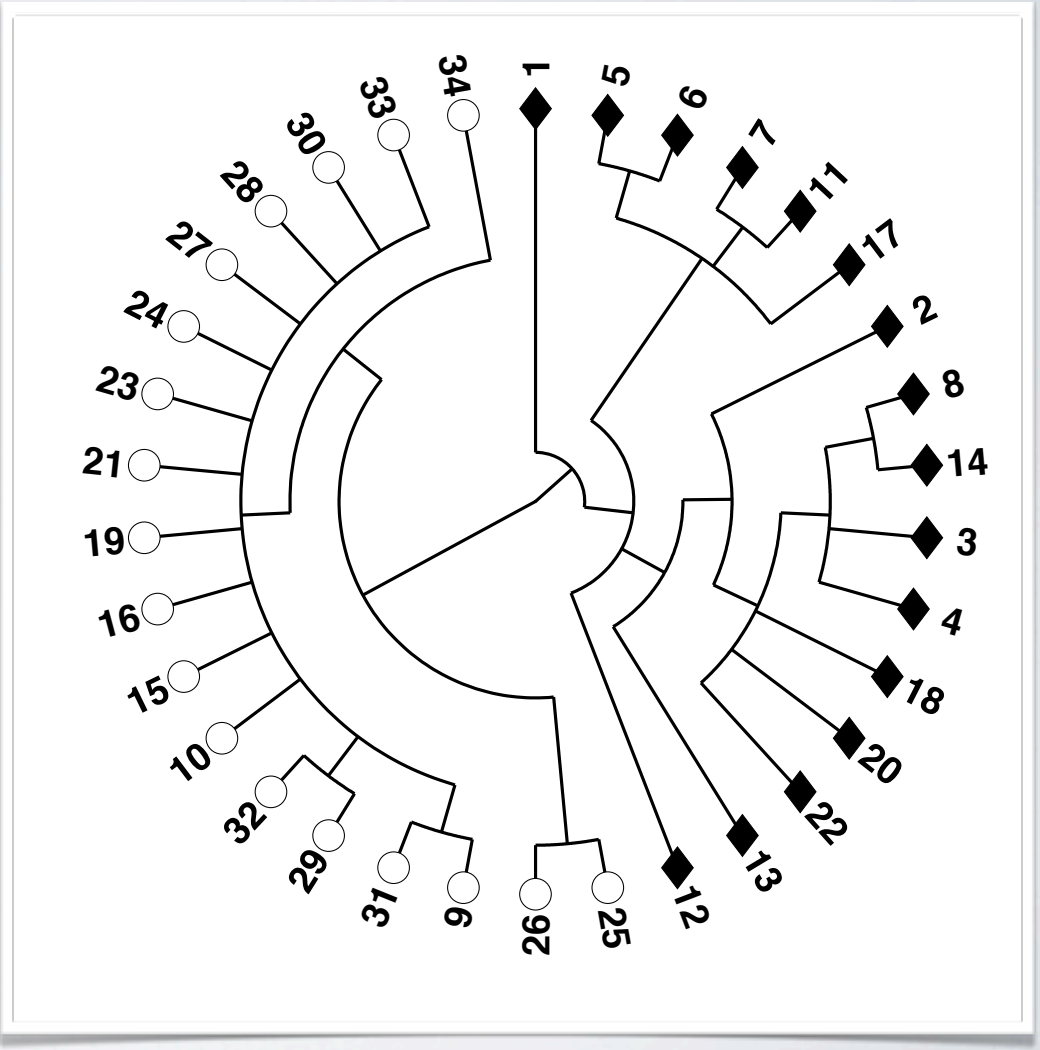


point and consensus hierarchies



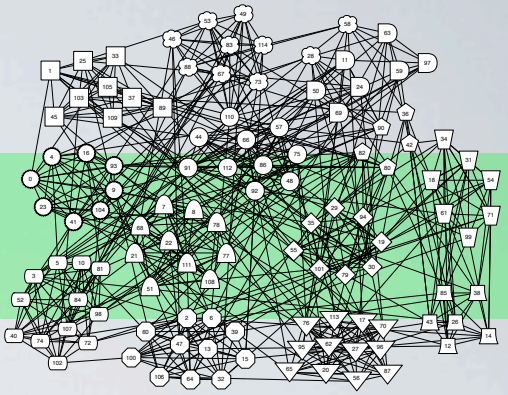


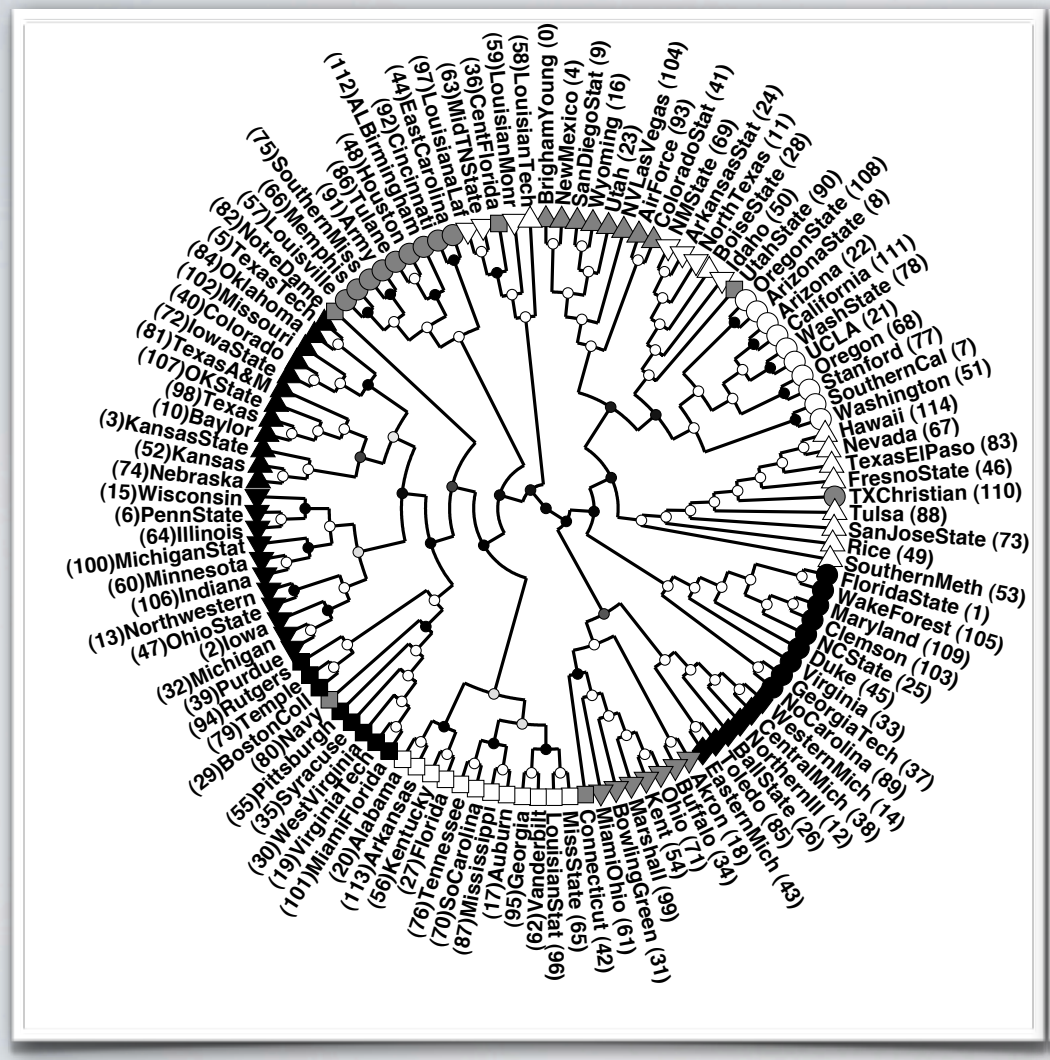
point estimate



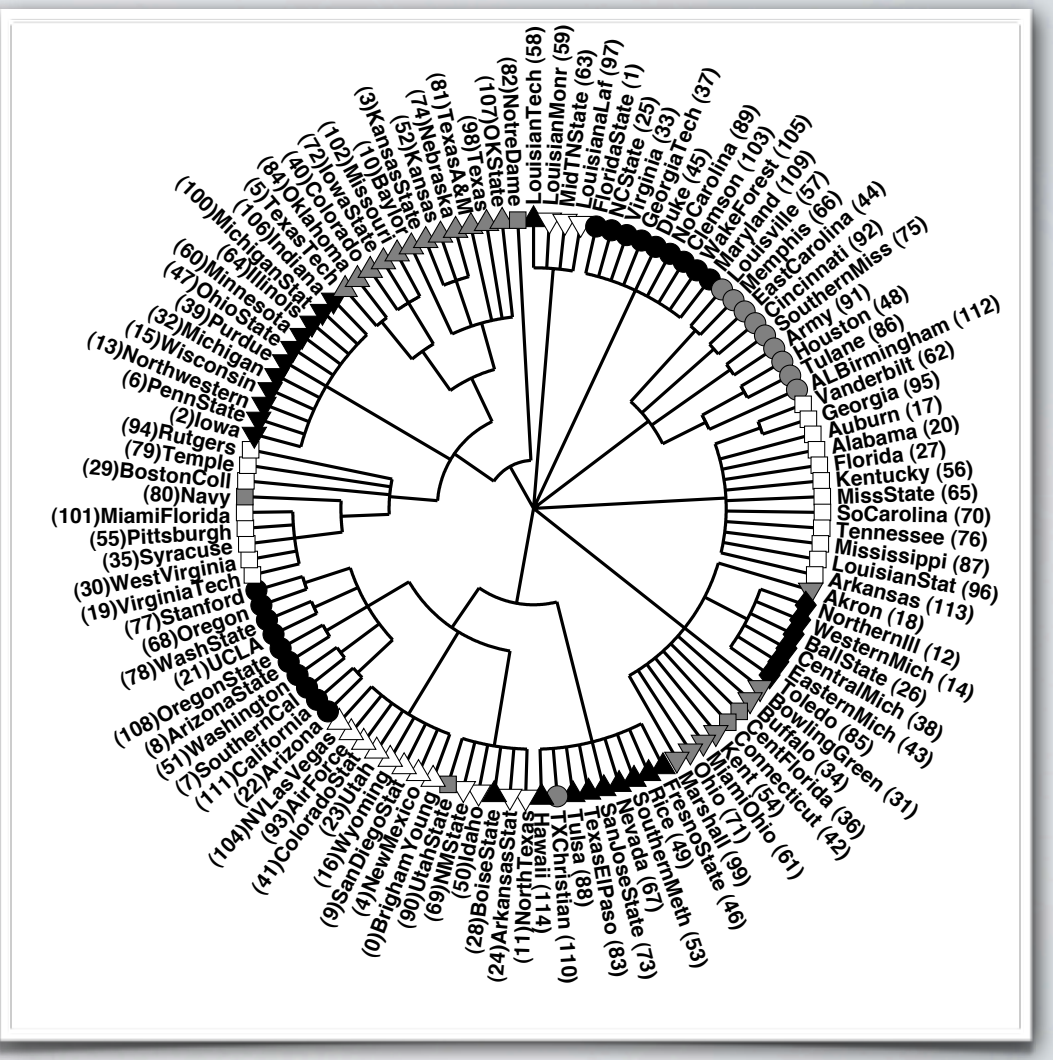
consensus hierarchy

point and consensus hierarchies





point estimate



consensus hierarchy

predicting missing links

many networks partially known, noisy

- social nets, foodwebs, protein interactions, etc.

use generative model to predict **missing links**

other approaches

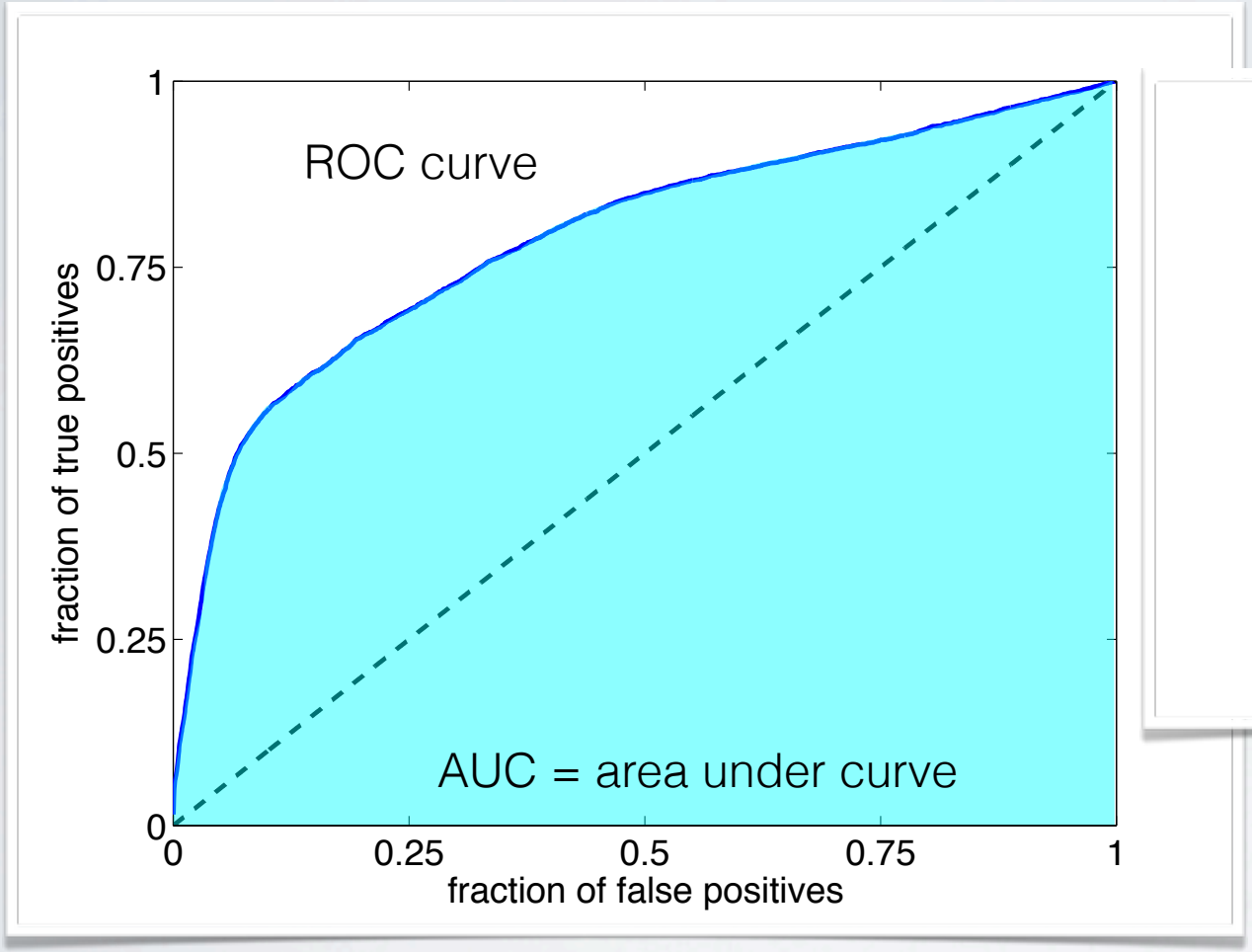
- Liben-Nowell & Kleinberg (2003)
- Goldberg & Roth (2003)
- Szilágyi et al. (2005)
- and now many others

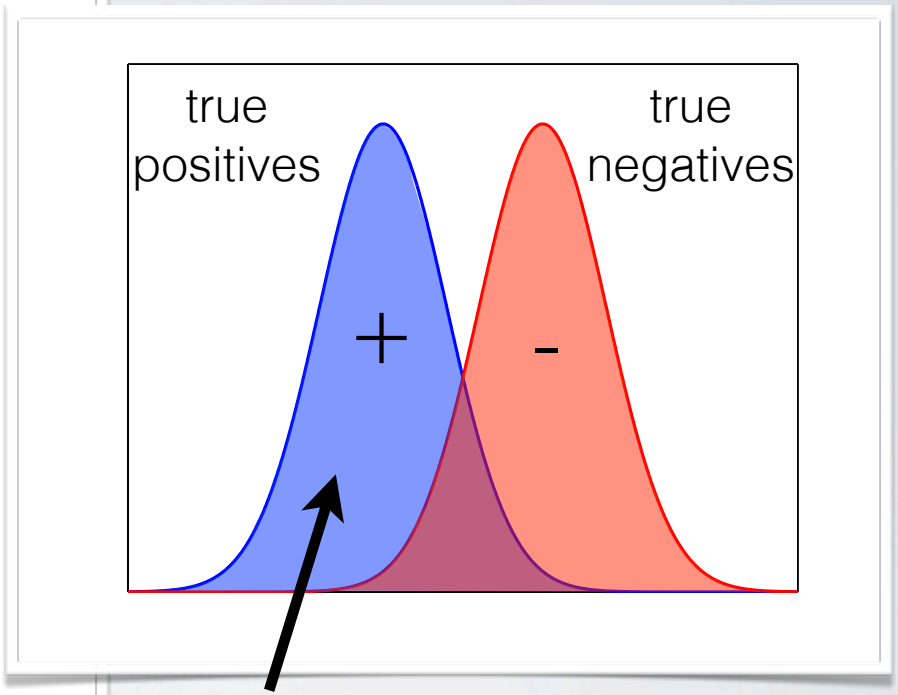
predicting missing links

- given incomplete graph G
- run MCMC to equilibrium
- then, over sampled $\mathcal{D}$, compute average $\langle p_r \rangle$ for links $(i, j) \notin G$
- predict links with high $\langle p_r \rangle$ values are missing

compute AUC via leave-*k*-out (edges) cross-validation

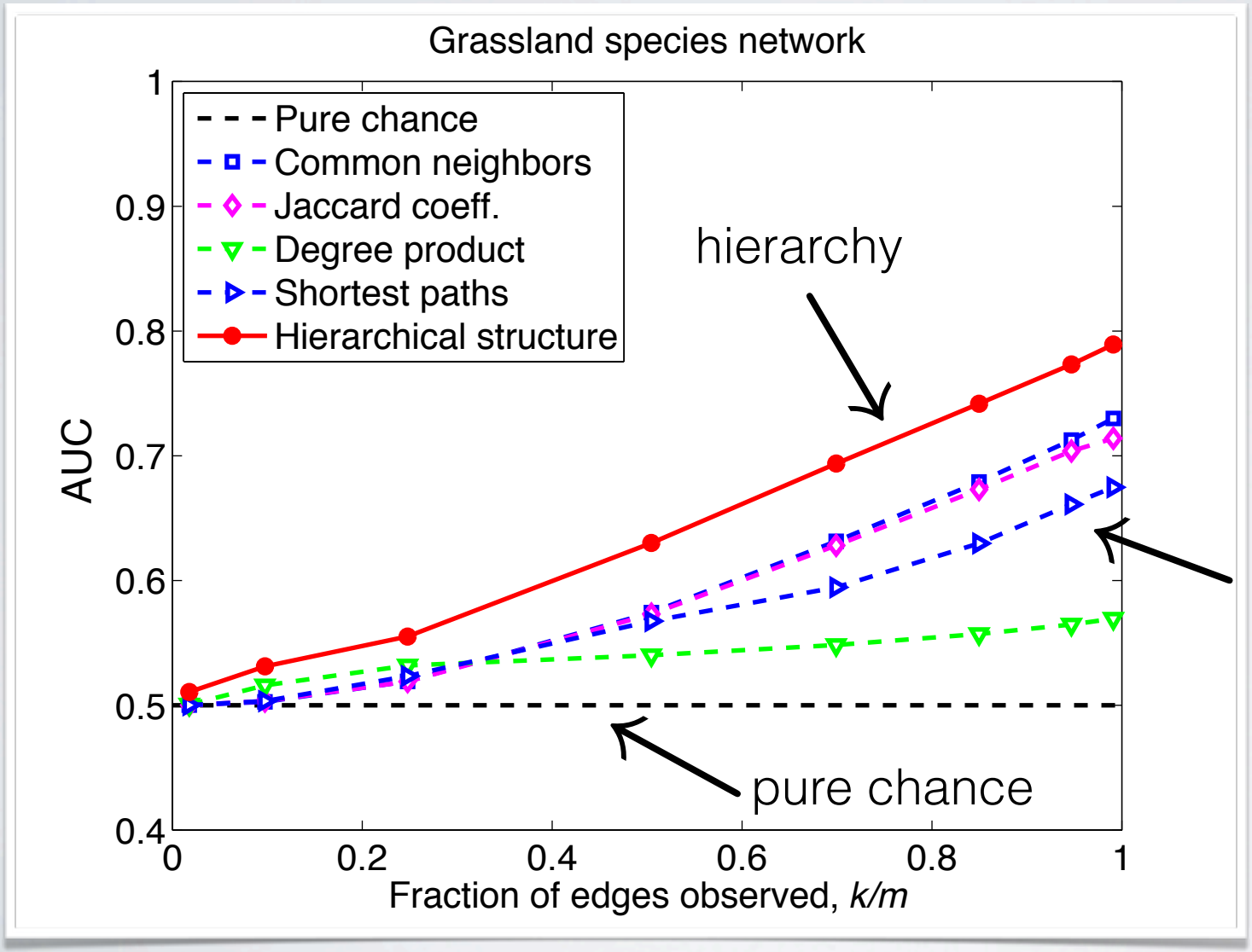
scoring the predictions

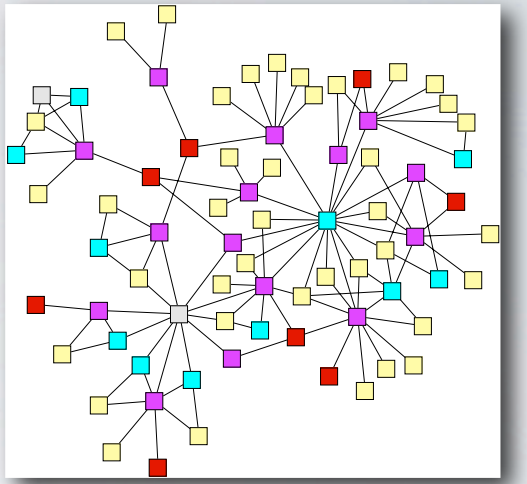




AUC =
Pr(distinguish
+ from -)

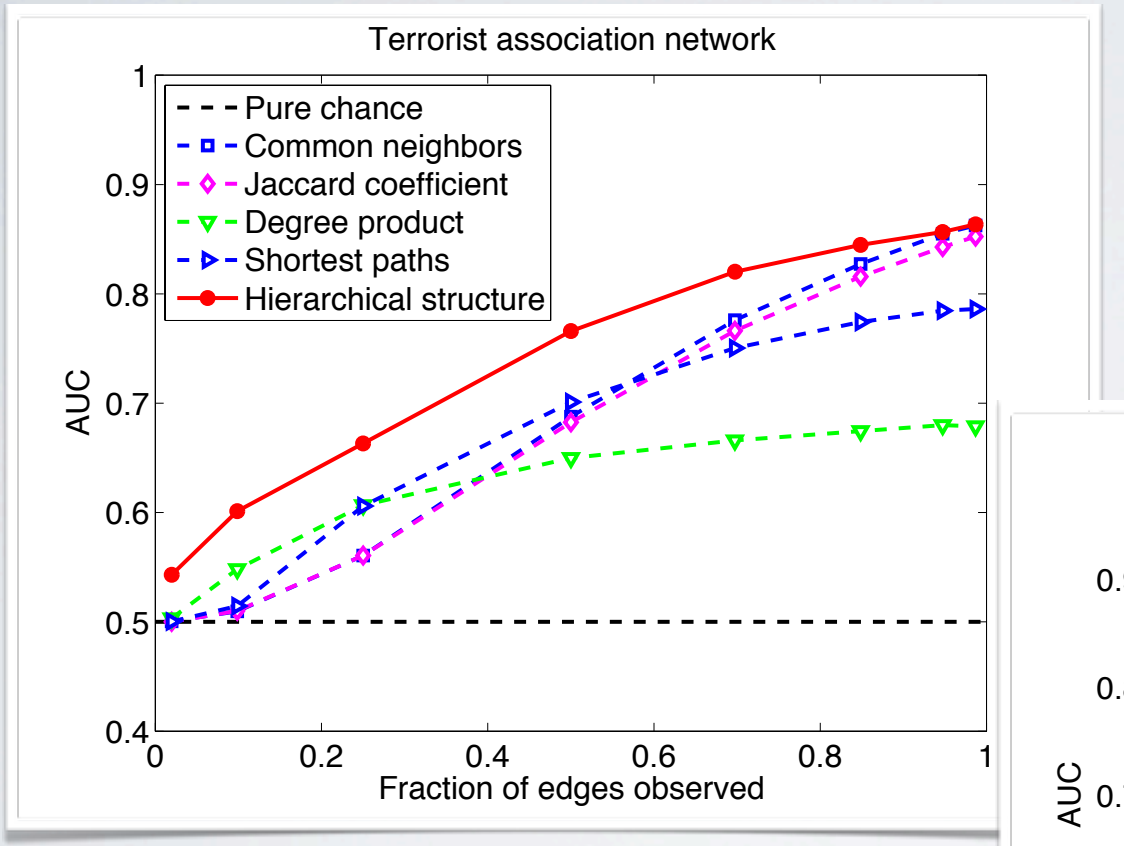
predicting missing links

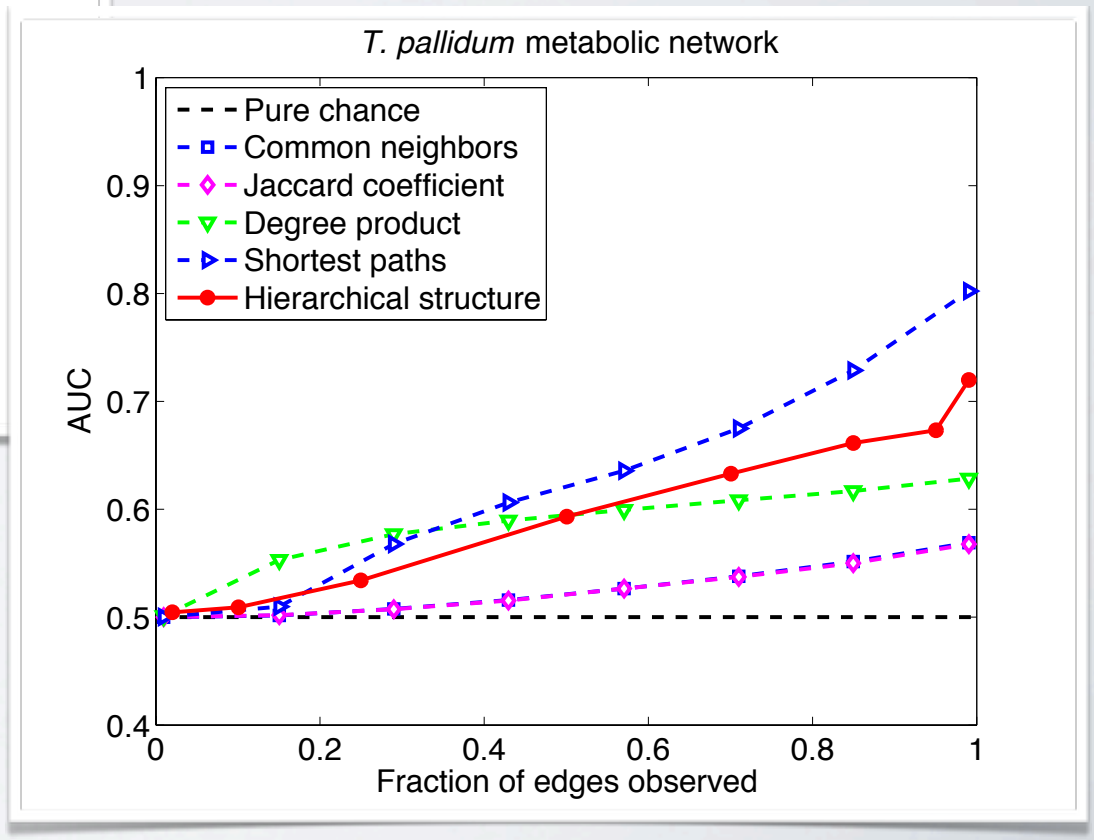




simple predictors

predicting missing links





final thoughts

- hierarchical organization
  - **non-random assembly of small scale structures**
  - naturally captures local topologies
  - effective multi-scale coarse-graining

- what's next?
  - scalability
  - modify attachment kernel to incl. add vertex/edge features
  - topological dynamics (e.g., fluxes, epidemics, signaling)
  - hierarchy formation mechanisms

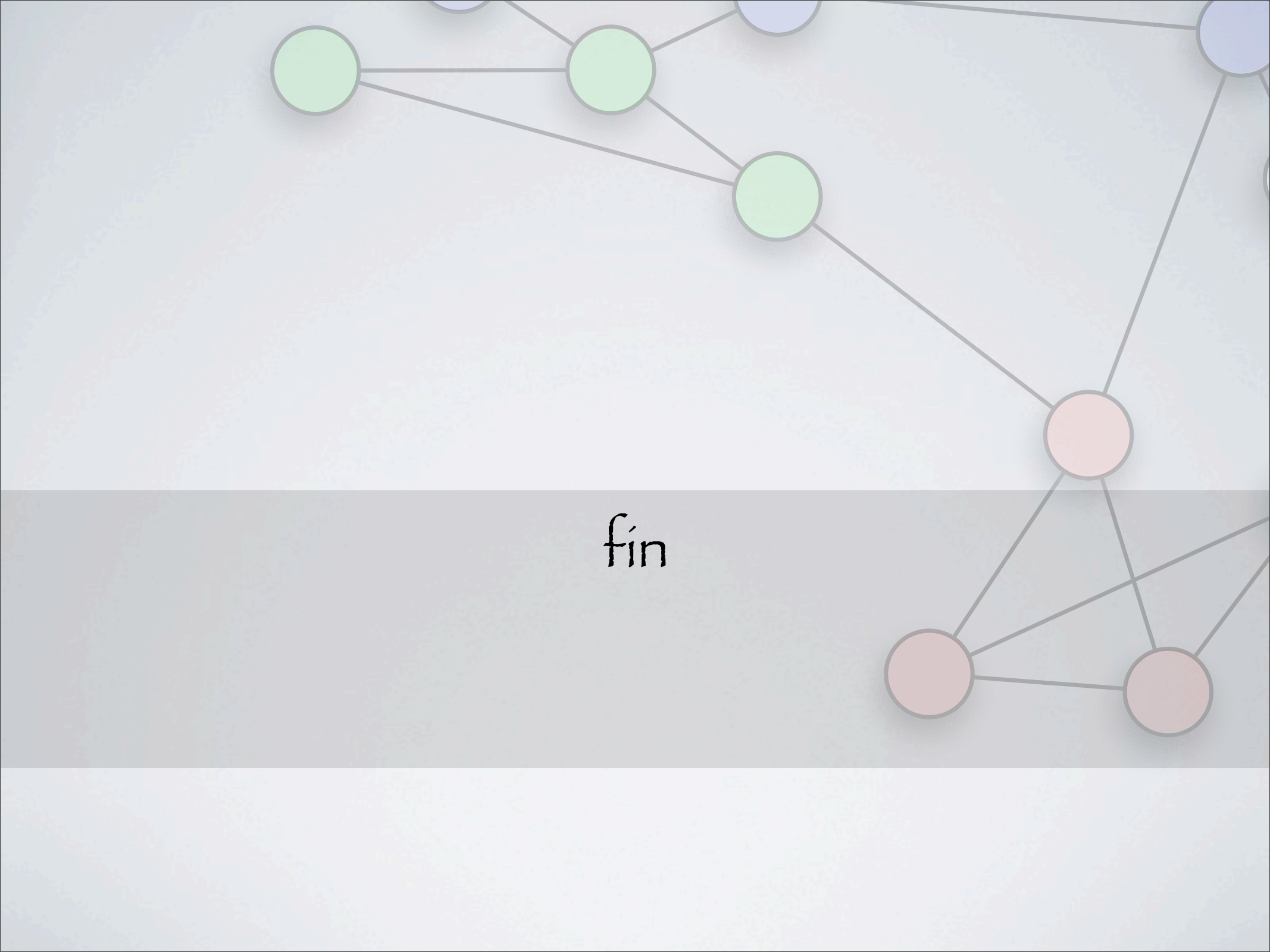
acknowledgements

joint work with:
Cris Moore (Santa Fe Institute)
Mark Newman (Michigan)

funding support:
National Science Foundation
James S. McDonnell Foundation

some references

- Yan, Zhu, Rouquier and Moore, "Active learning for hidden attributes in networks." *Proc. KDD* (2011)
- Park, Moore and Bader, "Dynamic Networks from Hierarchical Bayesian Graph Clustering." *PLoS ONE* **5**(1): e8118 (2010).
- Good, de Montjoye and Clauset, "Performance of modularity maximization in practical contextx." *Physical Review E* **81**, 046106 (2010)
- Clauset, Newman and Moore, "Hierarchical structure and the prediction of missing links in networks" *Nature* **453**, 98-101 (2008)
- Clauset, Moore and Newman, "Structural Inference of Hierarchies in Networks." *Proc. ICML* (2006)
- Clauset, Newman and Moore, "Finding community structure in very large networks." *Physical Review E* **70**, 066111 (2005)


fin